A General Theory of Goodness of Fit in Likelihood Fits

Rajendran Raja

Fermi National Accelerator Laboratory

Batavia, IL 60510

Abstract

Maximum likelihood fits to data can be performed using binned data and unbinned data. The likelihood fits in either case result in only the fitted quantities but not the goodness of fit. With binned data, one can obtain a measure of the goodness of fit by using the χ^2 method, after the maximum likelihood fitting is performed. With unbinned data, currently, the fitted parameters are obtained but no measure of goodness of fit is available. This remains, to date, an unsolved problem in statistics. By considering the transformation properties of likelihood functions with respect to change of variable, we conclude that the likelihood ratio of the theoretically predicted probability density to that of *the data density* is invariant under change of variable and provides the goodness of fit. We show how to apply this likelihood ratio for binned as well as unbinned likelihoods and show that even the χ^2 test is a special case of this general theory. In order to calculate errors in the fitted quantities, we use Bayes' theorem which then yields the surprising result that the quantity generally considered the Bayesian prior is an uninteresting constant and the resulting statistics is consistent with frequentist ideas.

Contents

Email address: raja@fnal.gov (Rajendran Raja).

1 Introduction

In particle physics as well as other branches of science, fitting theoretical models to data is a crucial end stage to the performance of experiments. Minimizing the χ^2 between theory and experiment is perhaps the most commonly used form of fitting, with data binned in histograms. Such fits yield not only the fitted parameters and errors on the fitted parameters but also a measure of the goodness of fit. Another common fitting method is the maximum likelihood method which can be performed on binned and unbinned data to obtain the best values of theoretical parameters. In the case of unbinned likelihood fitting, there is currently no measure of the goodness of fit. In this paper, we propose a solution to the problem, which by its nature works generally for both binned and unbinned likelihood fits. A general theory of goodness of fit in likelihood fits results.

In what follows, we will denote by the vector s, the theoretical parameters (s for "signal") and the vector c, the experimentally measured quantities or "configurations". For simplicity, we will illustrate the method where both s and c are one dimensional, though either or both can be multi-dimensional in practice. We thus define the theoretical model by the conditional probability density $P(c|s)$, defined as the probability of observing c given a value of s. The theoretical probability function obeys the normalization condition

$$\int P(c|s)dc = 1 \tag{1}$$

Then an unbinned maximum likelihood fit to data is obtained by maximizing the likelihood [1],

$$\mathcal{L} = \prod_{i=1}^{i=n} P(c_i|s) \tag{2}$$

where the likelihood is evaluated at the n observed data points $c_i, i = 1, n$. Such a fit will determine the maximum likelihood value s^* of the theoretical parameters, but will not tell us how good the fit is.

1.1 To show that $\mathcal{L}$ cannot be used as a goodness of fit variable

The goodness of fit variable must be invariant under a change of variable $c \rightarrow c'$. The value of the likelihood $\mathcal{L}$ at the maximum likelihood point does not furnish a goodness of fit, since the likelihood is not invariant under change of variable. This can be seen by observing that one can transform the variable set c to a variable set c' such that $P(c'|s^*)$ is uniformly distributed between 0 and 1. In one dimension, this is trivially done by the transformation function $c'(c)$ such that

$$c'(c) = \int_{c_1}^{c} P(t|s^*)dt \tag{3}$$

The variable c ranges from c_1 to c_2 and the probability function $P(c|s^*)$ normalizes to unity in this range. This implies that c' ranges from 0 to 1. Such a transformation is known as a hypercube transformation, in multi-dimensions. The transformed probability distribution in the variable c' is unity in this interval as can be seen by examining the Jacobian of the transformation $|\frac{\partial c'}{\partial c}|$

$$|\frac{\partial c'}{\partial c}| = P(c|s^*) \tag{4}$$

$$P(c'|s^*) = P(c|s^*)|\frac{\partial c}{\partial c'}| = 1 \tag{5}$$

Other datasets will yield different values of likelihood in the variable space c when the likelihood is computed with the original function $P(c|s^*)$. However, in hypercube space, the value of the likelihood is unity regardless of the dataset $c'_i, i = 1, n$, thus the likelihood $\mathcal{L}$ cannot furnish a goodness of fit by itself, since neither the likelihood, nor ratios of likelihoods computed using the same distribution $P(c|s^*)$ is invariant under variable transformations. The fundamental reason for this non-invariance is that only a single distribution, namely, $P(c|s^*)$ is being used to compute the goodness of fit.

To illustrate further, we use a concrete example of fitting a dataset using the maximum likelihood method as shown in Figure 1(a). The fitting is done in the range $c_1 < c < c_2$, where $c_1 = 1.0$ and $c_2 = 5.0$. The fitting function is

$$P(c|s) = \frac{\exp(-c/s)}{s(\exp(-c_1/s) - \exp(-c_2/s))} \tag{6}$$

which normalizes to unity in the range $c_1 < c < c_2$. The fitted dataset is shown as a full histogram. The dashed histogram shows a dataset that is a poor fit to the data and will produce a smaller value of $\mathcal{L}$ when fitted as a function of c. Figure 1(b) shows the same data in the hypercube space where the fitted function is flat as per the transformation given in equation 3. Both the datasets will produce a value of unity for $\mathcal{L}$ in this space implying an equally good fit in either case, which is obviously false. This clearly demonstrates that the likelihood by itself cannot provide a goodness of fit variable.

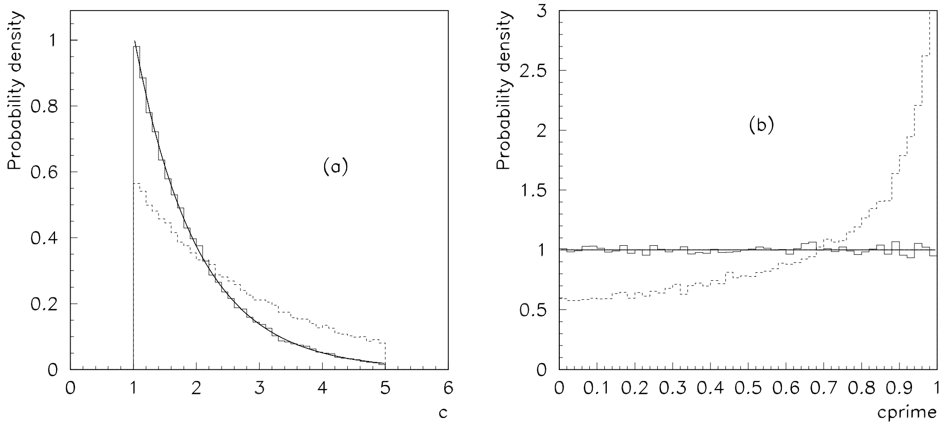

Fig. 1. (a) shows the fitting in the dataset space. The curve shows the fitted function. Superimposed is the fitted data, (full histogram, normalized to unity). The dashed histogram shows the different dataset which obviously does not fit to the fitted curve. (b) The same plot in hyperspace. the fitted function is flat by construction. Both the fitted data set (full histogram) and the dashed histogram will have the same value of likelihood $\mathcal{L}$ in this space which implies that $\mathcal{L}$ cannot be used as a goodness of fit variable.

2 Likelihood ratios

2.1 The concept of "data likelihood" derived from the pdf of the data

It is interesting to note that while using χ^2 as the goodness of fit technique for binned histograms, we use two distribution functions, namely the theoretical curve and the data. By binning the data, we are in effect estimating the probability density function of the data as the second distribution, in addition to the theoretical distribution specified by the theoretical curve. In likelihood language we define the probability density function (*pdf*) of the data as

$$P^{data}(c) = \lim_{n\to\infty} \frac{1}{n}\frac{dn}{dc} \tag{7}$$

which obeys the normalization condition

$$\int P^{data}(c)dc = 1 \tag{8}$$

When one is using binned likelihoods, the *pdf* of the data would be estimated by binning the events in a histogram and normalizing the sum of contents of all bins to unity. In the unbinned case, we will describe below a technique [2] on estimating $P^{data}(c)$ using Probability Density Estimators (PDE).

We can now define a likelihood ratio $\mathcal{L}_\mathcal{R}$ such that

$$\mathcal{L}_\mathcal{R} = \frac{\prod_{i=1}^{i=n} P(c_i|s)}{\prod_{i=1}^{i=n} P^{data}(c_i)} \equiv \frac{P(\vec{c_n}|s)}{P^{data}(\vec{c_n})} \tag{9}$$

where we have used the notation $\vec{c_n}$ to denote the dataset $c_i, i = 1, n$.

Since the n events $c_i, i = 1, n$ are independent, the probability of obtaining the dataset $\vec{c_n}$ is given by

$$P^{data}(\vec{c_n}) = \prod_{i=1}^{i=n} P^{data}(c_i) \tag{10}$$

The quantity $P^{data}(\vec{c_n})$ we name the "data likelihood" of the dataset $\vec{c_n}$ and the quantity $P(\vec{c_n}|s)$ as the "theory likelihood" of the dataset $\vec{c_n}$. We note that the "data likelihood" $P^{data}(\vec{c_n})$ may also be thought of as the probability density of of the" $n - object$" $\vec{c_n}$ which obeys the normalization condition

$$\int P^{data}(\vec{c_n})\ d\vec{c_n} = 1 \tag{11}$$

Let us now note that $\mathcal{L}_\mathcal{R}$ is invariant under a general variable transformation (not restricted to hypercube transformation) $c \to c'$, since

$$P(c'|s) = |\frac{\partial c}{\partial c'}|P(c|s) \tag{12}$$

$$P^{data}(c') = |\frac{\partial c}{\partial c'}|P^{data}(c) \tag{13}$$

$$\mathcal{L}'_{\mathcal{R}} = \mathcal{L}_{\mathcal{R}} \tag{14}$$

and the Jacobian of the transformation $|\frac{\partial c}{\partial c'}|$ cancels in the numerator and denominator in the ratio. This is an extremely important property of the likelihood ratio $\mathcal{L}_{\mathcal{R}}$ that qualifies it to be a goodness of fit variable. Since the denominator $P^{data}(\vec{c_n})$ is independent of the theoretical parameters s, both the likelihood ratio and the likelihood maximize at the same point s^*. The likelihood ratios for two different data sets $\vec{c_m}$ and $\vec{c_n}$ can be combined by multiplication as per

$${\mathcal{L}_{\mathcal{R}}}^{m+n} = {\mathcal{L}_{\mathcal{R}}}^{m} \times {\mathcal{L}_{\mathcal{R}}}^{n} \tag{15}$$

This rule follows from the definition of $\mathcal{L}_{\mathcal{R}}$ in equation 9. In practice, we will use the negative log-likelihood ratio $\mathcal{NLLR} = -log_e\mathcal{L}_R$ as the goodness of fit variable and minimize it. The multiplication rule of equation 15 results in an addition rule for $\mathcal{NLLR}$. The problem of finding the distribution of $\mathcal{NLLR}$ for a good fit then reduces to finding the distribution of $\mathcal{NLLR}$ in hyper-cube space for a variable that is uniformly distributed between zero and one, as in Figure 1(b). This is because $\mathcal{NLLR}$ is invariant under the transformation of variable. So all goodness of fit problems using likelihood ratios can be reduced to finding the distribution of $\mathcal{NLLR}$ for a variable that is uniformly distributed in hypercube space.

3 Normalizing the theoretical curve to the data

The method of maximum likelihood fits the shape of the theoretical distribution to the data distribution. The theoretical model obeys the normalization condition in equation 1 and the likelihood is evaluated at the number of observed data events n. There is no explicit mention of the theoretically expected number of events, which we denote by n_t. Later we will show how to incorporate a goodness of fit in the absolute normalization by making use of the binomial distribution and its limiting cases the Poisson and the normal distributions. We will begin by obtaining goodness of fit formulae for the case where we bin the data and fit the theoretical shape to the experimental distribution.

4 Binned Goodness of Fit

When one bins data in histograms and fits the theory shape to the data, one can fit by using either maximum likelihood or by minimizing χ^2. In either case, the goodness of fit is usually evaluated using χ^2. We now illustrate how the likelihood ratio defined in section 2 can be used to obtain a goodness of fit after the maximum likelihood fitting is done. In order to evaluate the likelihood ratio, one needs to evaluate the theory likelihood and the data likelihood for each value of c_i. For the binned histogram, we make the approximation of assuming that both these quantities are constant for all values of c_i in a given bin and evaluating each at the bin center. Let there be n_b bins and let the k^{th} bin contain n_k entries.

4.1 The multinomial distribution

The probability of obtaining the histogram is given by the multinomial distribution

$$P(histogram) = \frac{n!}{\prod_{k=1}^{k=n_b} n_k!} \prod_{k=1}^{k=n_b} P(c_k|s)^{n_k} \tag{16}$$

$$\sum_{k=1}^{k=n_b} n_k = n \tag{17}$$

4.2 Degeneracy of the distribution

The factor $\frac{n!}{\prod_{k=1}^{k=n_b} n_k!}$ denotes the number of ways n events can be partitioned to form the observed histogram, which we term the degeneracy $\mathcal{D}$ of the histogram. Each of the $\mathcal{D}$ histograms is identical to each other and possesses the same goodness of fit. We can then evaluate the goodness of fit for any one of the $\mathcal{D}$ degenerate histograms, the likelihood for which is given by

$$\mathcal{L} = \prod_{k=1}^{k=n_b} P(c_k|s)^{n_k} \tag{18}$$

and the likelihood ratio can be written as

$$\mathcal{L}_R = \prod_{k=1}^{k=n_b} \left(\frac{P(c_k|s)}{P^{data}(c_k)} \right)^{n_k} \tag{19}$$

The value of $\frac{P(c_k|s)}{P^{data}(c_k))}$ is raised to the power n_k in equation 19 results from the fact that there are n_k configurations c_i in the k^{th} bin and we are multiplying a constant ratio (at the bin center) over n_k configurations. If Δc_k is the bin width for the k^{th} bin, then the data likelihood can be approximated by

$$P^{data}(c_k) \approx \frac{n_k}{n \Delta c_k} \tag{20}$$

This obeys the normalization condition

$$\int P^{data}(c_k)dc_k \approx \sum_{k=1}^{k=n_b} \frac{n_k}{n\Delta c_k}\Delta c_k = 1. \tag{21}$$

The theoretical likelihood can be integrated over the bin to yield

$$P^{bin\,average}(c_k|s) = \frac{1}{\Delta c_k}\int\limits_{c=c_k-\Delta c_k/2}^{c=c_k+\Delta c_k/2} P(c|s)dc \tag{22}$$

This obeys the normalization condition

$$\sum_{k=1}^{k=n_b} P^{bin\,average}(c_k|s)\Delta c_k = 1 \tag{23}$$

Then the likelihood ratio can be written

$$\mathcal{L_R} = \prod_{k=1}^{k=n_b}\left(\frac{n\Delta c_k P^{bin\,average}(c_k|s)}{n_k}\right)^{n_k} \equiv \prod_{k=1}^{k=n_b}\left(\frac{T_k}{n_k}\right)^{n_k} \tag{24}$$

where $T_k \equiv n\Delta c_k P^{bin\,average}(c_k|s)$ is the theoretically expected number of events in the k^{th} bin obeying the normalization condition $\sum_k T_k = n$, as per equation 23. This likelihood ratio may be used to obtain a maximum likelihood fit as well as to obtain a goodness of fit. Note that the likelihood ratio is well-behaved even for empty bins where $n_k = 0$, since $n_k^{n_k}$ is unity for such cases.

Note that the negative log-likelihood ratio $\mathcal{NLLR}$ resulting from equation 24 yields

$$\mathcal{NLLR} = \sum_{k=1}^{k=n_b} n_k\, log_e\,(\frac{n_k}{T_k}) \tag{25}$$

which is the same result as derived by Baker and Cousins [3] for the multinomial case where normalization is preserved between theory and experiment.

We have derived the result using very different arguments (than Baker and Cousins) for the denominator of the likelihood ratio, namely it is the value of the *data pdf* at the bin center as a result of the general theory developed here.

If we are reluctant to work out (for reasons of computing speed) the integral in equation 22 for each bin at each step of the fitting process, then we can approximate it by the bin center values

$$P^{bin\,average}(c_k|s) \approx \frac{P(c_k|s)}{\sum_k P(c_k|s)\;\Delta c_k} \tag{26}$$

This then obeys the normalization equation 23 and the expression in equation 25 for $\mathcal{NLLR}$ can be used generally.

4.3 To Show that the Binned Negative Log-Likelihood Ratio Approaches a χ^2 Distribution for Large n

Let the difference between n_k, the observed number of events and T_k the theoretical number of events be denoted by $\lambda_k = n_k - T_k$. Then $\sum_k \lambda_k = 0$, by virtue of the normalization conditions. Then the binned negative log likelihood ratio $\mathcal{NLLR}$ can be written

$$\mathcal{NLLR} = -log_e\;\mathcal{L}_R = -\sum_{k=1}^{k=n_b} n_k\;log_e\left(1 - \frac{\lambda_k}{n_k}\right) \tag{27}$$

This can be expanded in powers of λ_k/n_k as

$$\mathcal{NLLR} = -log_e\;\mathcal{L}_R = \sum_{k=1}^{k=n_b} n_k\left(\frac{\lambda_k}{n_k} + \frac{1}{2}(\frac{\lambda_k}{n_k})^2 + \frac{1}{3}(\frac{\lambda_k}{n_k})^3 + \frac{1}{4}(\frac{\lambda_k}{n_k})^4\cdots\right) \tag{28}$$

$$= \sum_{k=1}^{k=n_b} \frac{1}{2}(\frac{\lambda_k^2}{n_k}) + \frac{1}{3}(\frac{\lambda_k^3}{n_k^2}) + \frac{1}{4}(\frac{\lambda_k^4}{n_k^3})\cdots \tag{29}$$

As $n \to \infty$, the individual bin contents become normally distributed about their expected value T_k with variance $\sigma_k^2 = n_k(1 - n_k/n) \approx n_k$ for $n_k << n$. This is true for all cases (named the *null hypothesis*) where the data and theory fit each other. Then we can write $\chi_k^2 = \lambda_k/n_k$ and

$$\mathcal{NLLR} = \sum_{k=1}^{k=n_b} \frac{1}{2}\chi_k^2 + \frac{1}{3}\frac{\lambda_k^3}{\sigma_k^4} + \frac{1}{4}\frac{\lambda_k^4}{\sigma_k^6} \cdots \tag{30}$$

For large n, $\lambda_k \approx \sqrt{n}_k$ and the higher order terms may be neglected yielding

$$\mathcal{NLLR} \to \sum_{k=1}^{k=n_b} \frac{1}{2}\chi_k^2 \; when \, n \to \infty \tag{31}$$

The expected value of the $\mathcal{NLLR}$ can then be written

$$E(\mathcal{NLLR}) = \sum_{k=1}^{k=n_b} \frac{1}{2}E(\chi_k^2) + \frac{1}{3}\frac{\mu_3}{\sigma_k^4} + \frac{1}{4}\frac{\mu_4}{\sigma_k^6} + \frac{1}{5}\frac{\mu_5}{\sigma_k^8} + \frac{1}{6}\frac{\mu_6}{\sigma_k^{10}} \cdots \tag{32}$$

where $\mu_3, \mu_4, \cdots$ are the $3^{rd}, 4^{th} \cdots$ moments of the normal distribution about the mean. Since the normal distribution is symmetric about the mean, all the odd moments $(\mu_3, \mu_5 \cdots)$ are zero. The even moments of the normal distribution (for integer l) are given by the formula

$$\mu_{2l} = 1.3.5 \cdots (2l - 1)\sigma^{2l} \tag{33}$$

This yields

$$E(\mathcal{NLLR}) = \sum_{k=1}^{k=n_b} \frac{1}{2}E(\chi_k^2) + \frac{3}{4}\frac{\sigma_k^4}{\sigma_k^6} + \frac{15}{6}\frac{\sigma_k^8}{\sigma_k^{10}} \cdots \tag{34}$$

All the remaining terms tend to zero as $1/n_k (= 1/\sigma_k^2)$ as $n_k \to \infty$ leading to

$$E(\mathcal{NLLR}) = \sum_{k=1}^{k=n_b} \frac{1}{2}E(\chi_k^2) = \frac{n_b}{2} \tag{35}$$

$$E(\mathcal{L}_R) = \exp(-n_b/2) \tag{36}$$

The number of degrees of freedom for $\mathcal{NLLR}$ would be $n_b - 1$, due to the normalization condition $\sum_k n_k = n$.

4.4 Normalizing theory and experiment and the problem of Goodness of fit for the Poisson distribution

As we have pointed out, maximum likelihood fitting only fits the shape of the theoretical distribution to the experimental data. This is due to the normalization condition of equation 1. However, if we employ a binomial distribution and define the first bin as containing the number of observed events n with theoretical expectation of n_t events, and the second bin to contain the number of unobserved events in N tries, then one can employ the formula in equation 24 with $n_b = 2$ to obtain the likelihood ratio.

$$\mathcal{L}_R = \left(\frac{n_t}{n}\right)^n \left(\frac{N - n_t}{N - n}\right)^{N-n} = \left(\frac{n_t}{n}\right)^n \left(\frac{1 - n_t/N}{1 - n/N}\right)^{N-n} \tag{37}$$

We now take the Poissonian limit of $N \rightarrow \infty$ with n_t and n finite and the above likelihood ratio becomes

$$\mathcal{L}_R = e^{-(n_t - n)} \left(\frac{n_t}{n}\right)^n \tag{38}$$

where we have employed the relations $(N - n) \rightarrow N$ and $(1 - x/N)^N \rightarrow e^{-x}$ as $N \rightarrow \infty$.

Equation 38 provides the goodness of fit likelihood ratio for all Poissonian problems where n_t events are expected and n are observed. We can now multiply this Poissonian $\mathcal{L}_R$ with equation 24 to produce the likelihood ratio for a general binned likelihood problem where the normalization for theory and experiment vary.

$$\mathcal{L}_R = e^{-(n_t - n)} \left(\frac{n_t}{n}\right)^n \prod_{k=1}^{k=n_b} \left(\frac{T_k}{n_k}\right)^{n_k} = e^{-(n_t-n)} \prod_{k=1}^{k=n_b} \left(\frac{T'_k}{n_k}\right)^{n_k} \tag{39}$$

where we have defined $T'_k = n_t T_k / n$ and $\sum T'_k = n_t$. With this redefinition, we obtain the $\mathcal{NLLR}$ for the multinomial with theoretical normalization differing from the experimental one as

$$\mathcal{NLLR} = \sum_{k=1}^{k=n_b} T'_k - n_k + n_k \, log_e(\frac{n_k}{T'_k}) \tag{40}$$

This is same as the "Poissonian result" of Baker and Cousins [3] again derived using very different arguments for the denominator of the likelihood ratio.

4.5 The Gaussian limit of the binomial

The Poissonian result is useful when n_t and n are relatively small numbers ($<\approx 25$). When we have larger number of events, then the Gaussian approximation is more relevant. We have already shown that (equation 29) that in a multinomial, the negative log likelihood ratio can be approximated by

$$\mathcal{NLLR} = \sum_{k=1}^{k=n_b} \frac{1}{2} \left(\frac{\lambda_k^2}{n_k}\right) \tag{41}$$

We apply this to the binomial with $n_b = 2$, $n_1 = n$, and $n_2 = N - n$ and $\lambda_1 = -\lambda_2 = n - n_t$. Then

$$\mathcal{NLLR} = \frac{\lambda_2^2}{2} \left(\frac{1}{n_1} + \frac{1}{n_2}\right) = \frac{\lambda_2^2}{2} \left(\frac{1}{(1 - n/N)(n/N)N}\right) \tag{42}$$

$$\approx \frac{\lambda_2^2}{2} \left(\frac{1}{Npq}\right) = \frac{(n - n_t)^2}{2\sigma^2} \tag{43}$$

where $p = n_t/N \approx n/N$ is the probability of an event appearing in the first bin and $q = 1 - p$ and $\sigma^2 = Npq$ is the variance of the bin contents of the first bin. We now let $N \to \infty$, $n \to \infty$ and $N >> n$. In this case, the variance can be approximated by n and we have the Gaussian case with $\mathcal{NLLR} = (n - n_t)^2/2n)$. This $\mathcal{NLLR}$ can be added to the one resulting from the maximum likelihood shape fitting to get an overall goodness of fit.

We must emphasize once again that the method of maximum likelihood always fits theoretical shapes to experimental data. We have been able to circumvent this restriction by using the device of the binomial distribution where the observed events n are in the first bin and the total number of events in the distribution N refer to the "number of tries" and the second bin consists of the $N-n$ events that failed to appear in the experiment. The binomial distribution is special in this regard since once we specify the properties of the first bin, the second bin is completely specified and anti-correlated with the first bin. The number of tries is unknown, but we set it to infinity in two different limits as discussed resulting in the Poisson and the Gaussian likelihood ratios.

4.6 To show that χ^2 is also the negative logarithm of a likelihood ratio

The most commonly used method for goodness of fit is the χ^2 test of Karl Pearson, which is used even when the quantities being fitted are not events but measurements with error bars. We show here that the χ^2 measure is also twice the negative logarithm of a Gaussian likelihood *ratio* rather than the negative logarithm of a Gaussian likelihood, as is the popular misconception. Consider a binned histogram where the contents in the k^{th} bin is noted by c_k and the theoretical expectation of this bin is s_k. The standard error of the observed variable c_k is known to be σ_k. Then, one can write

$$P(c_k|s_k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(c_k - s_k)^2}{2\sigma_k^2}\right) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{\chi_k^2}{2}\right) \tag{44}$$

This leads to

$$-\log_e\left(P(c_k|s_k)\right) = \frac{\chi_k^2}{2} + \log_e(\sqrt{2\pi}\sigma_k) \tag{45}$$

From the above expression, people are mistakenly led to conclude that χ^2 is equivalent to twice the negative log-likelihood. This ignores the term $\log_e(\sqrt{2\pi}\sigma_k)$ in the above equation, which varies from bin to bin. In order to work out the likelihood ratio, we need to estimate the data density $P(c_k)$ at each measurement. The data points are distributed as a Gaussian with standard deviation σ_k. The best estimate of the mean of the Gaussian from the data alone is c_k. This leads to

$$P(c_k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(c_k - c_k)^2}{2\sigma_k^2}\right) = \frac{1}{\sqrt{2\pi}\sigma_k} \tag{46}$$

yielding the likelihood ratio

$${\mathcal{L}_{\mathcal{R}}}^k = \frac{P(c_k|s_k)}{P(c_k)} = \exp\left(-\frac{(s_k - c_k)^2}{2\sigma_k^2}\right) = \exp(-\frac{\chi_k^2}{2}) \tag{47}$$

The overall likelihood ratio is given by

$$\mathcal{L}_R = \prod_{k=1}^{k=n_b} \mathcal{L}_R^k \tag{48}$$

leading to

$$\chi^2 = 2\ log_e\left(\mathcal{L}_R\right) = \sum_{k=1}^{k=n_b} \chi_k^2 \tag{49}$$

i.e. χ^2 is equal to twice the negative log-likelihood ratio and not the negative log-likelihood!.

5 Unbinned Goodness of Fit

Very often the data are not plentiful enough to bin adequately and it is more efficient to perform an unbinned likelihood fit. Presently, a goodness of fit method does not exist for unbinned likelihood fits. Using the formalism developed above, we present a solution. After the unbinned likelihood fit is performed by maximizing the likelihood in equation 2 one needs to work out the *data likelihood* $P^{data}(\vec{c_n})$ in order to evaluate the likelihood ratio and the goodness of fit. We employ the technique of Probability Density Estimators $(PDE's)$, also known as Kernel Density Estimators [2] $(KDE's)$ to do this. The *pdf* $P^{data}(c)$ is approximated by

$$P^{data}(c) \approx PDE(c) = \frac{1}{n}\sum_{i=1}^{i=n} \mathcal{G}(c - c_i) \tag{50}$$

where a Kernel function $\mathcal{G}(c - c_i)$ is centered around each data point c_i, is so defined that it normalizes to unity. The choice of the Kernel function can vary depending on the problem. A popular kernel is the Gaussian defined in the multi-dimensional case as

$$\mathcal{G}(c) = \frac{1}{(\sqrt{2\pi}h)^d\sqrt{(det(E))}} exp(\frac{-H^{\alpha\beta}c^\alpha c^\beta}{2h^2}) \tag{51}$$

where E is the error matrix of the data defined as

$$E^{\alpha,\beta} =< c^\alpha c^\beta > - < c^\alpha >< c^\beta > \tag{52}$$

and the <> implies average over the n events, and d is the number of dimensions. The Hessian matrix H is defined as the inverse of E and the repeated indices imply summing over. The parameter h is a "smoothing parameter", which has[5] a suggested optimal value $h \propto n^{-1/(d+4)}$, that satisfies the asymptotic condition

$$\mathcal{G}_\infty(c - c_i) \equiv \lim_{n\to\infty} \mathcal{G}(c - c_i) = \delta(c - c_i) \tag{53}$$

The parameter h will depend on the local number density and will have to be adjusted as a function of the local density to obtain good representation of the data by the PDE. Our proposal for the goodness of fit in unbinned likelihood fits is thus the likelihood ratio

$$\mathcal{L}_\mathcal{R} = \frac{P(\vec{c_n}|s)}{P^{data}(\vec{c_n})} \approx \frac{P(\vec{c_n}|s)}{P^{PDE}(\vec{c_n})} \tag{54}$$

evaluated at the maximum likelihood point s^*.

6 An illustrative example

We consider a simple one-dimensional case where the data is an exponential distribution, say decay times of a radioactive isotope. The theoretical prediction is given by

$$P(c|s) = \frac{1}{s}\exp(-\frac{c}{s}) \tag{55}$$

We have chosen an exponential with $s = 1.0$ for this example. The Gaussian Kernel for the PDE would be given by

$$\mathcal{G}(c) = \frac{1}{(\sqrt{2\pi}\sigma h)} \exp(-\frac{c^2}{2\sigma^2 h^2}) \tag{56}$$

where the variance σ of the exponential is numerically equal to s. To begin with, we chose a constant value for the smoothing parameter, which for 1000 events generated is calculated to be 0.125. Figure 2 shows the generated events, the theoretical curve $P(c|s)$ and the PDE curve $P(c)$ normalized to the number of events. The PDE fails to reproduce the data near the origin due to the boundary effect, whereby the Gaussian probabilities for events close to the origin spill over to negative values of c. This lost probability would be compensated by events on the exponential distribution with negative c if they existed. In our case, this presents a drawback for the PDE method, which we will remedy later in the paper using PDE definitions on the hypercube and periodic boundary conditions. For the time being, we will confine our example to values of $c > 1.0$ to avoid the boundary effect.

In order to test the goodness of fit capabilities of the likelihood ratio $\mathcal{L}_{\mathcal{R}}$, we superimpose a Gaussian on the exponential and try and fit the data by a simple exponential. Figure 3 shows the "data" with 1000 events generated as an exponential in the fiducial range $1.0 < c < 5.0$. Superimposed on it is a Gaussian of 500 events. More events in the exponential are generated in the interval $0.0 < c < 1.0$ to avoid the boundary effect at the fiducial boundary at c=1.0. Since the number density varies significantly, we have had to introduce a method of iteratively determining the smoothing factor as a function of c as described in [4]. With this modification in the PDE, one gets a good description of the behavior of the data by the PDE as shown in Figure 3.

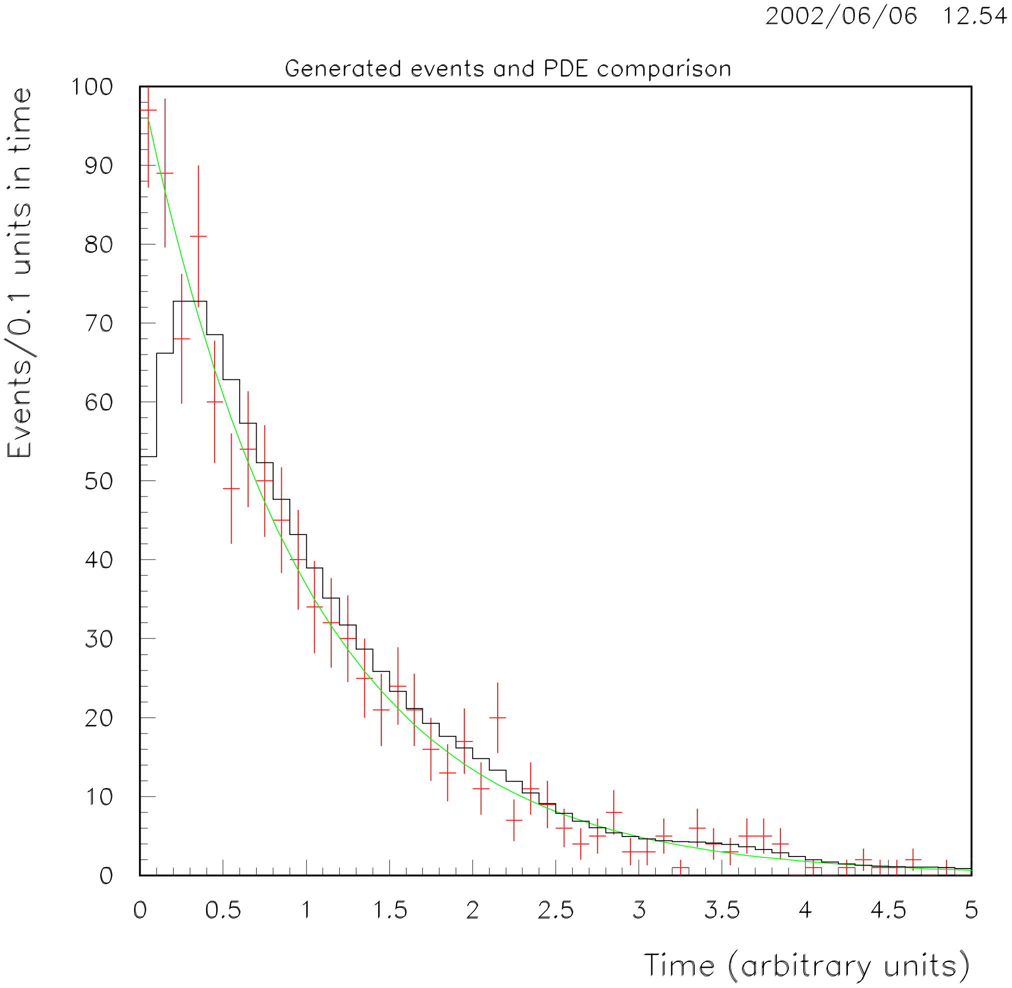

Fig. 2. Figure shows the histogram (with errors) of generated events. Superimposed is the theoretical curve $P(c|s)$ and the PDE estimator (solid) histogram with no errors.

We now vary the number of events in the Gaussian and obtain the value of the negative log likelihood ratio $\mathcal{NLLR}$ as a function of the strength of the Gaussian. Table 1 summarizes the results. The number of standard deviations the unbinned likelihood fit is from what is expected is determined empirically by plotting the value of $\mathcal{NLLR}$ for a large number of fits where no Gaussian is superimposed (i.e. the null hypothesis) and determining the mean and RMS of this distribution and using these to estimate the number of σ's the observed

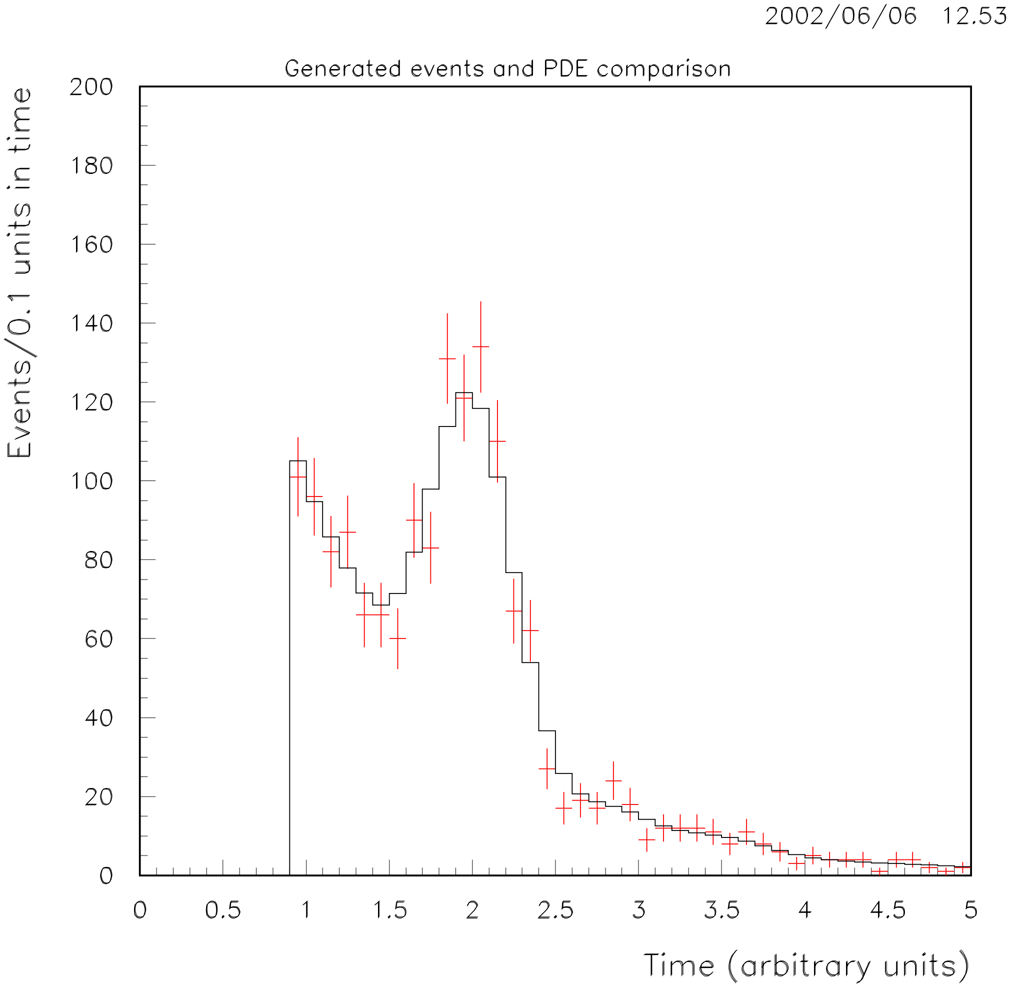

Fig. 3. Figure shows the histogram (with errors) of 1000 events in the fiducial interval $1.0 < c < 5.0$ generated as an exponential with decay constant s=1.0 with a superimposed Gaussian of 500 events centered at c=2.0 and width=0.2. The PDE estimator is the (solid) histogram with no errors.

$\mathcal{NLLR}$ is from the null case. Table 1 also gives the results of a binned fit on the same "data". It can be seen that the unbinned fit gives a 3σ discrimination when the number of Gaussian events is 85, where as the binned fit gives a χ^2/ndf of 42/39 for the same case.

Table 1

Goodness of fit results from unbinned likelihood and binned likelihood fits for various data samples. The negative values for the number of standard deviations in some of the examples is due to statistical fluctuation.

Number of	Unbinned fit	Unbinned fit	Binned fit χ^2
Gaussian events	$\mathcal{NLLR}$	$N\sigma$	39 d.o.f.
500	189.	103	304
250	58.6	31	125
100	11.6	4.9	48
85	8.2	3.0	42
75	6.3	1.9	38
50	2.55	-0.14	30
0	0.44	-1.33	24

Figure 4 shows the variation of -log $P(\vec{c_n}|s)$ and -log $P^{PDE}(\vec{c_n})$ for an ensemble of 500 experiments each with the number of events $n = 1000$ in the exponential and no events in the Gaussian (null hypothesis). It can be seen that -log $P(\vec{c_n}|s)$ and -log $P^{PDE}(\vec{c_n})$ are correlated with each other and the difference between the two (-log $\mathcal{NLLR}$) is a much narrower distribution than either and provides the goodness of fit discrimination.

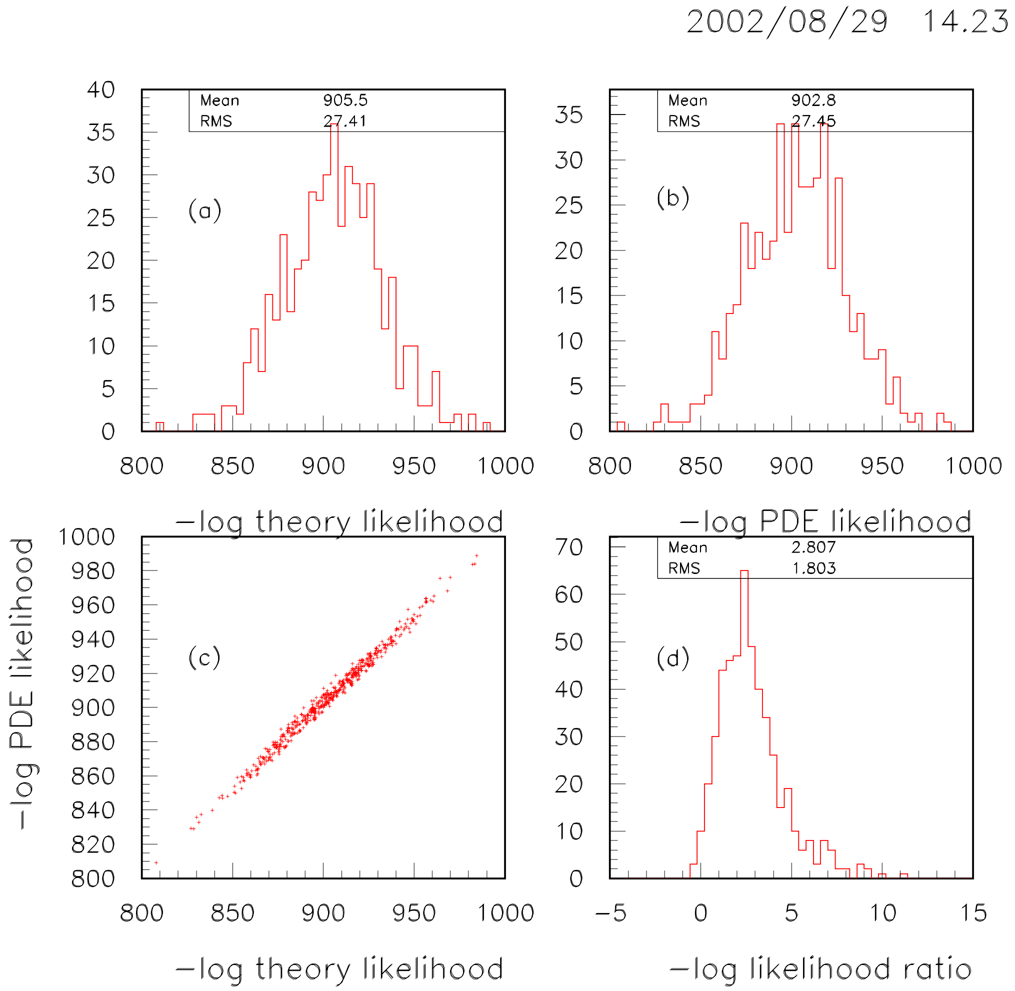

Fig. 4. (a) shows the distribution of the negative log-likelihood -$log_e(P(\vec{c_n}|s))$ for an ensemble of experiments where data and experiment are expected to fit. (b) Shows the negative log PDE likelihood -$log_e(P(\vec{c_n}))$ for the same data (c) Shows the correlation between the two and (d) Shows the negative log-likelihood ratio $\mathcal{NLLR}$ that is obtained by subtracting (b) from (a) on an event by event basis.

6.1 Improving the PDE

The PDE technique we have used so far suffers from two drawbacks; firstly, the smoothing parameter has to be iteratively adjusted significantly over the full range of the variable c, since the distribution $P(c|s)$ changes significantly over that range; and secondly, there are boundary effects at c=0 as shown in figure 2. Both these flaws are remedied if we define the PDE in hypercube space. After we find the maximum likelihood point s^*, for which the PDE is not needed, we transform the variable $c \rightarrow c'$, such that the distribution $P(c'|s^*)$ is flat and $0 < c' < 1$. The hypercube transformation can be made even if c is multi-dimensional by initially going to a set of variables that are uncorrelated and then making the hypercube transformation. The transformation can be such that any interval in c space maps on to the interval $(0, 1)$ in hypercube space.

6.2 Periodic Boundary Conditions

We solve the boundary problem by imposing periodicity in the hypercube. In the one dimensional case, we imagine three "hypercubes", each identical to the other on the real axis in the intervals $(-1, 0)$, $(0, 1)$ and $(1, 2)$. The hypercube of interest is the one in the interval $(0, 1)$. When the probability from an event kernel leaks outside the boundary $(0, 1)$, we continue the kernel to the next hypercube. Since the hypercubes are identical, this implies the kernel re-appearing in the middle hypercube but from the opposite boundary. Put mathematically, the kernel is defined such that

$$\mathcal{G}(c' - c'_i) = \mathcal{G}(c' - c'_i - 1);\ c' > 1 \tag{57}$$

$$\mathcal{G}(c' - c'_i) = \mathcal{G}(c' - c'_i + 1);\ c' < 0 \tag{58}$$

Although a Gaussian Kernel will work on the hypercube, the natural kernel to use considering the shape of the distribution in hypercube space (it is flat for a good fit), would be the "boxcar function" $\mathcal{G}(c')$.

$$\mathcal{G}(c') = \frac{1}{h};\ |c'| < \frac{h}{2} \tag{59}$$

$$\mathcal{G}(c') = 0;\ |c'| > \frac{h}{2} \tag{60}$$

This kernel would be subject to the periodic boundary conditions given above, which further ensure that every configuration in hypercube space is treated exactly as every other configuration irrespective of its co-ordinate. The parameter h is a smoothing parameter which needs to be chosen with some care. However, since the theory distribution is flat in hypercube space, the smoothing parameter may not need to be iteratively determined over hypercube space to the extent that data distribution is similar to the theory distribution. Even if iteration is used, the variation in h in hypercube space is likely to be much smaller.

Figure 5 shows the distribution of the $\mathcal{NLLR}$ for the null hypothesis for an ensemble of 500 experiments each with 1000 events as a function of the smoothing factor h. It can be seen that the distribution narrows considerably as the smoothing factor increases. We choose an operating value of 0.2 for h and study the dependence of the $\mathcal{NLLR}$ as a function of the number of events ranging from 100 to 1000 events, as shown in figure 6. The dependence on the number of events is seen to be weak, indicating good behavior. The PDE thus arrived computed with h=0.2 can be transformed from the hypercube space to c space and will reproduce data smoothly and with no edge effects. We note that it is also easier to arrive at an analytic theory of $\mathcal{NLLR}$ with the choice of this simple kernel.

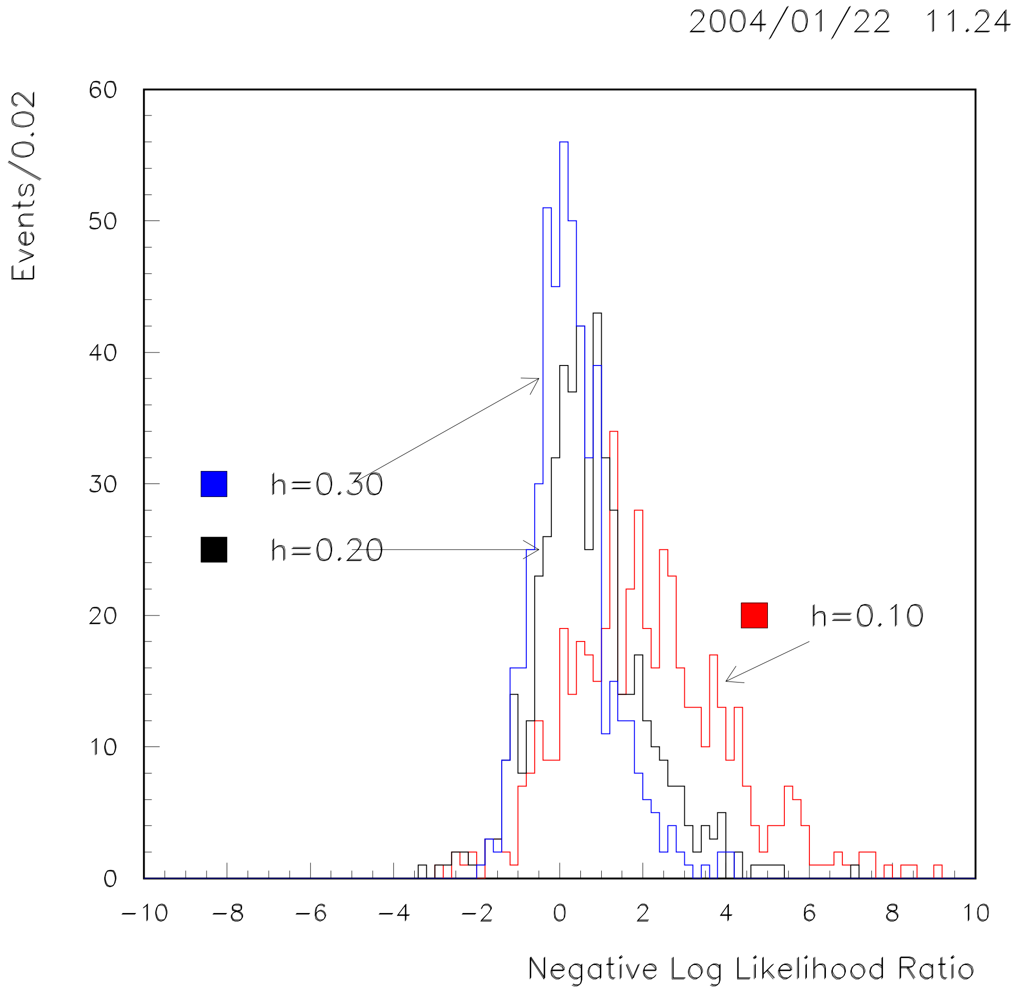

Fig. 5. The distribution of the negative log likelihood ratio $\mathcal{NLLR}$ for the null hypothesis for an ensemble of 500 experiments each with 1000 events, as a function of the smoothing factor h=0.1, 0.2 and 0.3

7 The distribution of the goodness of fit variable

Of all the goodness of fit variables we have studied above, for both binned and unbinned likelihood fits, the χ^2 variable is the most studied and has an analytic theory associated with its distribution. This is used to set a p-value for the goodness of fit, defined as the probablity to exceed the observed value χ^2 based on its analyic distribution. In the absence of an analytic theory, it is

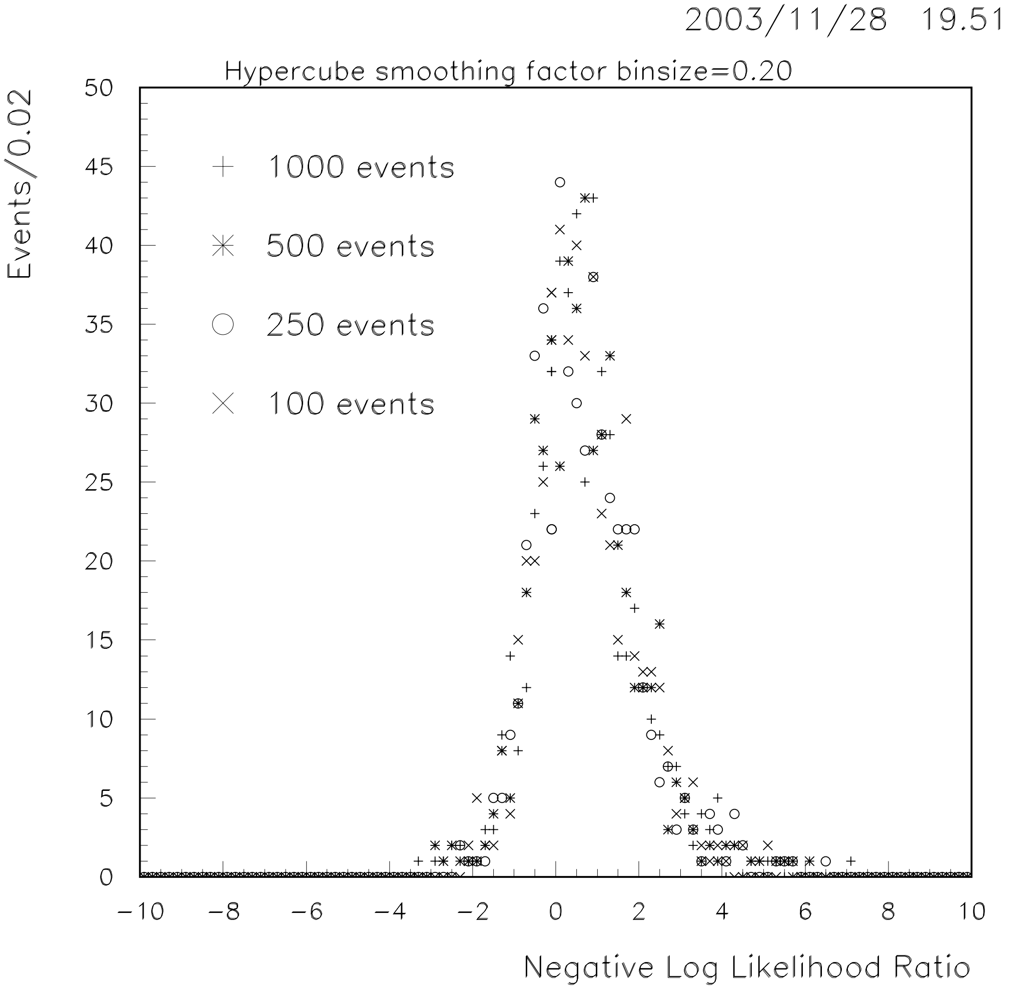

Fig. 6. The distribution of the negative log likelihood ratio $\mathcal{NLLR}$ for the null hypothesis for an ensemble of 500 experiments each with the smoothing factor h=0.2, as a function of the number of events

possible to use Monte Carlo methods to obtain the distribution of the goodness of fit variable for the hypothesis being tested and to numerically obtain the p-value.

8 Calculation of fitted errors

After the fitting is done and the goodness of fit is evaluated, one needs to work out the errors on the fitted quantities. One needs to calculate the posterior density $P(s|\vec{c_n})$, which carries information not only about the maximum likelihood point s^*, from a single experiment, but how such a measurement is likely to fluctuate if we repeat the experiment.

8.1 The concept of the pdf of a fixed parameter

Before we begin the error calculation, we would like to define precisely a few concepts. The theoretical parameter s is a fixed but unknown constant. What do we mean by its probability density function? First, determine the maximum likelihood value s^* form a single dataset $\vec{c_n}$. Repeat this procedure for an ensemble of such datasets. We define $\mathcal{P}_n(s)$ as the probability density function of the parameter s, the distribution of s^* that we would obtain from such an infinite ensemble of datasets. We employ the subscript n to note the possible dependence of the *pdf* on the number of elements n in each of the datasets in the ensemble.

8.2 The true value of the parameter s

The true value of the parameter s is defined to be that value of s at which the maximum of the *pdf* $\mathcal{P}_n(s)$ occurs. Let us remember that $\mathcal{P}_n(s)$ has an infinite number of similar datasets $\vec{c_n}$ contributing to it and hence this is just a statement of the experiments being unbiased.

8.3 The unknowability of $\mathcal{P}_n(s)$

Since the true value of s can never be determined to infinite precision, and the true value is the abscissa for which the *pdf* $\mathcal{P}_n(s)$ is the maximum, it follows that the function $\mathcal{P}_n(s)$ is unknowable. We cannot associate an abscissa to the function $\mathcal{P}_n(s)$ and hence the function cannot be "anchored" to the s axis. We thus call this function the "unknown concomitant", to distinguish it from a Bayesian prior.

8.4 The posterior density $P(s|\vec{c_n})$

In order to determine the error on the fitted parameter s, we need to determine the posterior density $P(s|\vec{c_n})$. The maximum likelihood fit yields the maximum likelihood value s^* given $\vec{c_n}$. We postulate that there is additional information in a single dataset $\vec{c_n}$ to yield an estimate of the distribution of s^* from an ensemble of such datasets. That information is expressed in the posterior density $P(s|\vec{c_n})$.

We would like to determine this function $P(s|\vec{c_n})$ using Bayes' theorem. Since Bayes' theorem is central to the argument, we give a simple and intuitively compelling derivation of it for two continuous variables c, s.

8.5 Derivation of Bayes' theorem equations

Consider a joint probability distribution $P(s, c)$ in variables s, c. For the sake of simplicity, we will take both s and c to be one-dimensional. The arguments being made are general enough to easily change them into multi-dimensional variables. Figure 7 shows geometrically the two dimensional space of s and c. We plot s as the ordinate and c as the abscissa. At this stage s and c are two general variables. Then,

$$\int\int P(s,c)dsdc = 1 \tag{61}$$

We define the single variable probabilities $P(c)$ and $P(s)$ as

$$P(c) = \int P(s,c)ds \tag{62}$$

$$P(s) = \int P(s,c)dc \tag{63}$$

$P(c)$ is the probability density of c irrespective of the value of s and $P(s)$ is the probability density of s irrespective of the value of c. It follows from equation 61 that

$$\int P(s)ds = 1 \tag{64}$$

and

$$\int P(c)dc = 1 \tag{65}$$

We define a conditional probability $P(c|s)$ as the probability of observing c given s. It is thus, the joint probability $P(s,c)$ along the slice AB (s=constant) in figure 7, appropriately normalized to unity. *i.e*,

$$P(c|s) = \frac{P(s,c)}{\int P(s,c)dc} \tag{66}$$

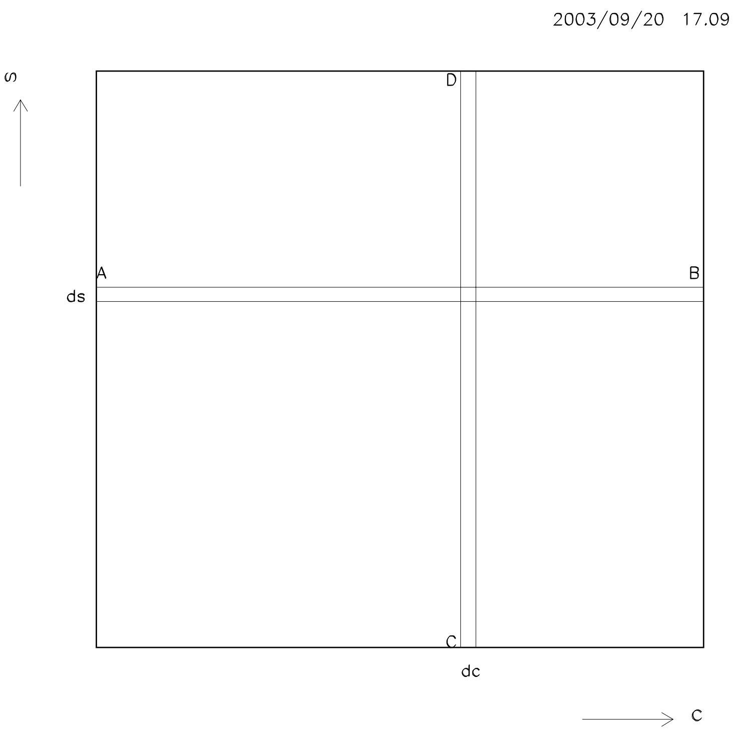

Fig. 7. Joint probability distribution in the variables s, c. Conditional probabilities are computed along the slices AB(s=constant) and CD(c= constant).

where the denominator in the above equation ensures that $\int P(c|s)dc = 1$. Therefore, (using equation 63)

$$P(c|s) = \frac{P(s,c)}{P(s)} \tag{67}$$

By symmetrical arguments (integrations along the slice CD), we show that the conditional probability $P(s|c)$ is given by

$$P(s|c) = \frac{P(s,c)}{P(c)} \tag{68}$$

leading to the joint probability equation

$$P(s,c) = P(c|s)P(s) = P(s|c)P(c) \tag{69}$$

which is sometimes written in a more familiar form known as Bayes' theorem [6] as

$$P(s|c) = \frac{P(c|s)P(s)}{P(c)} \tag{70}$$

It is a general theorem in statistics, which we have derived using intuitive geometrically explicit arguments. By substituting the expression for $P(s, c)$ in equation 67 in equation 62 we get the equation

$$P(c) = \int P(c|s)P(s)ds \tag{71}$$

and by substituting the expression for $P(s, c)$ in equation 68 in equation 63 we get the equation

$$P(s) = \int P(s|c)P(c)dc \tag{72}$$

These complete the Bayes' theorem equations. Note also that the joint probability equation 69 can be written in a form a likelihood ratio $\mathcal{L}_R$

$$\mathcal{L}_R = \frac{P(s|c)}{P(s)} = \frac{P(c|s)}{P(c)} \tag{73}$$

The quantity $\mathcal{L}_R$ equation 73 is invariant under change of variables $c \rightarrow c'$ and $s \rightarrow s'$, since the Jacobian of the transformation $|\frac{\partial c'}{\partial c}|$ divides out in the numerator and the denominator for the right hand side of the equation 73 for the ratio of probability densities in $\frac{P(c|s)}{P(c)}$. Similarly the ratio is invariant under the transformation variable s in the LHS of the equation. These invariances are essential in the use of the ratio $\mathcal{L}_R$ as a goodness-of-fit variable.

We can then extend the derivation given above to derive Bayes' theorem equations for the dataset $\vec{c_n}$.

$$P(s,\vec{c_n}) = P(\vec{c_n}|s)\mathcal{P}_n(s) = P(s|\vec{c_n})P^{data}(\vec{c_n}) \tag{74}$$

$$P^{data}(\vec{c_n}) = \int P(\vec{c_n}|s)\mathcal{P}_n(s)ds \tag{75}$$

$$\mathcal{P}_n(s) = \int P(s|\vec{c_n})P^{data}(\vec{c_n})d\vec{c_n} \tag{76}$$

Let us note that the above derivation of Bayes' theorem treats the variables c and s symmetrically. $P(c)$ and $P(s)$ are projections of the joint probability $P(s,c)$ on the c and s axes respectively. Neither $P(c)$ nor $P(s)$ is a prior in the Bayesian sense.

8.6 Determination of the Posterior Density $P(s|\vec{c_n})$

The joint probability density $P(s,\vec{c_n})$ of the parameter s and the data $\vec{c_n}$ is given by

$$P^{data}(s,\vec{c_n}) = P(s|\vec{c_n})P^{data}(\vec{c_n}) \tag{77}$$

where we use the superscript ^{data} to distinguish the joint probability $P^{data}(s,\vec{c_n})$ as having come from using the data *pdf*. If we now integrate the above equation over all possible datasets $\vec{c_n}$, we get the expression for (using equation 76) $\mathcal{P}_n(s)$.

$$\mathcal{P}_n(s) = \int P^{data}(s,\vec{c_n})d\vec{c_n} = \int P(s|\vec{c_n})P^{data}(\vec{c_n})d\vec{c_n} \tag{78}$$

Equation 78 states that in order to obtain the *pdf* of the parameter s, one needs to add together the conditional probabilities $P(s|\vec{c_n})$ over an ensemble of events, each such distribution weighted by the "data likelihood" $P^{data}(\vec{c_n})$. At this stage of the discussion, the function $P^{data}(s|\vec{c_n})$ is unknown. However, it is important to note that equation 78 enables us to write down an expression for the *pdf* of s, given the posterior density $P(s|\vec{c_n})$ and the key concept of the "data likelihood" $P^{data}(\vec{c_n})$ we have introduced, motivated by goodness of fit considerations.

8.7 The error bootstrap

The error on the fitted parameter s^* will be related to the width of the posterior density $P(s|\vec{c_n})$ that we are trying to compute. It is also related to our ignorance of the value of s_T and our inability to anchor the distribution $\mathcal{P}_n(s)$. At this stage, we have worked out $\mathcal{L}_{\mathcal{R}}(s)$ as a function of s and have evaluated the maximum likelihood value s^* of s. We can choose an arbitrary value of s and evaluate the goodness of fit at that value using the likelihood ratio. When we do this, we are in fact hypothesizing that s_T, the true value, is at this value of s. The function $\mathcal{L}_R(s)$ then gives us a way of evaluating the goodness of fit of the hypothesis as we change s. Let us now take an arbitrary value of s and hypothesize that that is the true value. Then consistent with our hypothesis, we must insist that the distribution $\mathcal{P}_n(s)$ is moved so that the maximum value of the distribution (i.e. s_T) is at the current value of s.

Then the theoretical estimate for the joint probability $P^{theory}(s, \vec{c_n})$ is given by the product of the probability density of the *pdf* of s at the true value of s, namely $\mathcal{P}_n(s_T)$, and the theoretical likelihood $P(c_n|s)$ evaluated at the true value, which by our hypothesis is s.

$$P^{theor}(s, \vec{c_n}) = P^{theor}(\vec{c_n}|s)\mathcal{P}_n(s_T) \tag{79}$$

The joint probability $P(s, \vec{c_n})$ is a joint distribution of the theoretical parameter s and data $\vec{c_n}$. The two ways of evaluating this (from the theoretical end and the data end) must yield the same result, for consistency. This is equivalent to equating $P^{data}(s, \vec{c_n})$ and $P^{theor}(s, \vec{c_n})$. This gives the equation

$$P(s|\vec{c_n})P^{data}(\vec{c_n}) = P^{theor}(\vec{c_n}|s)\mathcal{P}_n(s_T) \tag{80}$$

which is a form of Bayes' theorem, but with two $pdf's$ (theory and data).

Rearranging equation 80, one gets

$$P(s|\vec{c_n}) = \mathcal{L}_{\mathcal{R}}(s)\mathcal{P}_n(s_T) = \frac{P^{theor}(\vec{c_n}|s)}{P^{data}(\vec{c_n})}\mathcal{P}_n(s_T) \tag{81}$$

To reiterate, when one varies s in equation 81, one makes the hypothesis that $s = s_T$. As one changes s, a new hypothesis is being tested that is mutually exclusive from the previous one, since the true value can only be at one location. So as one changes s, one is free to move the *distribution* $\mathcal{P}_n(s)$ so that s_T is at the value of s being tested. This implies that $\mathcal{P}_n(s_T)$ does not change as one changes s and is a constant *wrt* s, which we can now write as α_n. Figure 8 illustrates these points graphically. Thus $\mathcal{P}_n(s_T)$ in our equations is a number, not a function. We have thus "bootstrapped" the error. On the one hand, $P(s|\vec{c_n})$ gives us an estimate of the spread in the measurements of s^* from an ensemble of datasets $\vec{c_n}$, based on one such data set. From the theoretical end, the error in s^* is expressed in the uncertainty on where to put s_T. We have connected these two uncertainties using Bayes' theorem and hypothesis testing. We can now solve for $P(s|\vec{c_n})$ as shown below.

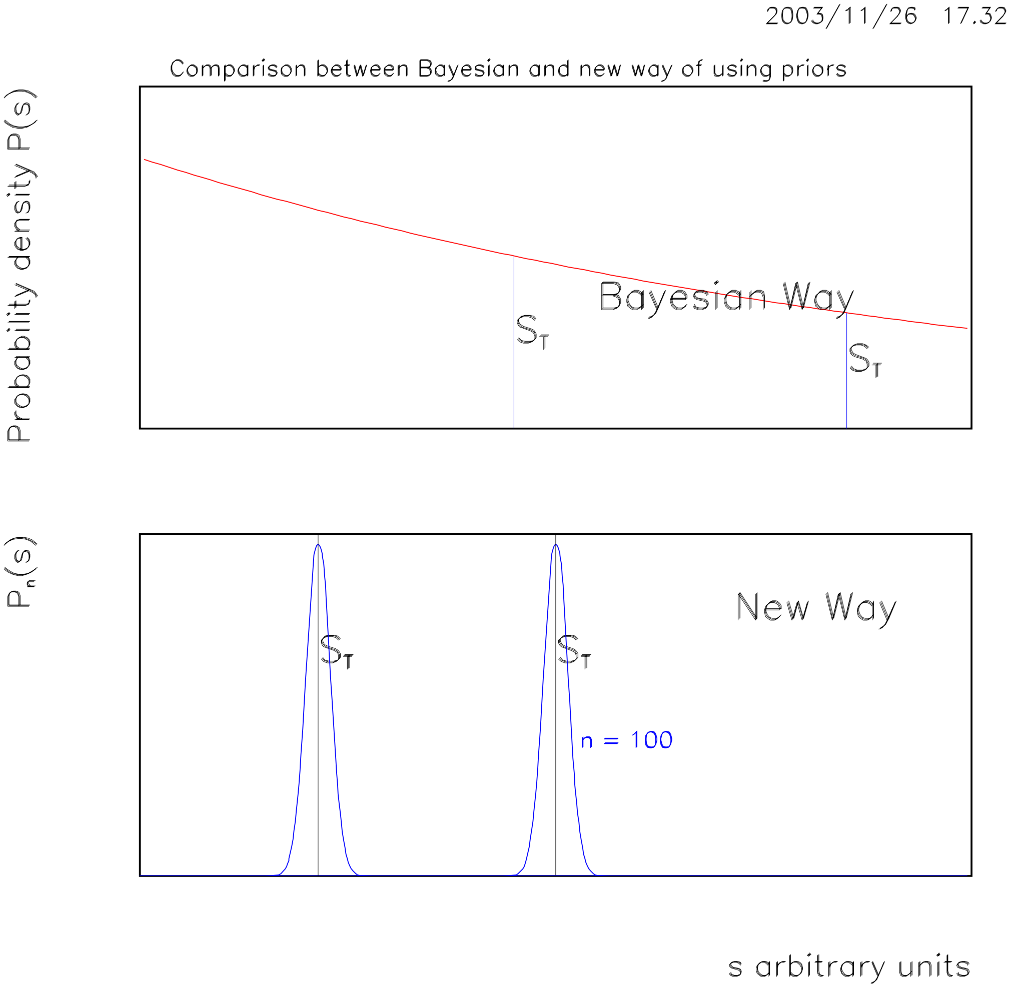

Fig. 8. Comparison of the usage of Bayesian priors with the new method. In the upper figure, illustrating the Bayesian method, an unknown distribution is guessed at by the user based on "degrees of belief" and the value of the Bayesian prior changes as the variable s changes. In the lower figure, an "unknown concomitant" distribution is used whose shape depends on the statistics. In the case of no bias, this distribution peaks at the true value of s. As we change s, we change our hypothesis as to where the true value of s lies, and the distribution shifts with s as explained in the text. The value of the distribution at the true value is thus independent of s.

8.8 New form of equations

Equation 81 can now be re-written

$$P(s|\vec{c_n}) = \frac{P(\vec{c_n}|s)\alpha_n}{P^{data}(\vec{c_n})} \tag{82}$$

Since $P(s|\vec{c_n})$ must normalize to unity, one gets for α_n,

$$\alpha_n = \frac{P^{data}(\vec{c_n})}{\int P(\vec{c_n}|s)ds} = \frac{1}{\int \mathcal{L}_{\mathcal{R}}(s)\ ds} \tag{83}$$

We have thus determined α_n, the value of the "unknown concomitant" at the true value s_T using our data set c_n. This is our *measurement* of α_n and different datasets will give different values of α_n, in other words α_n will have a sampling distribution with an expected value and standard deviation.

Note that it is only possible to write down an expression for α_n dimensionally when a likelihood ratio $\mathcal{L}_{\mathcal{R}}$ is available. This then leads to

$$P(s|\vec{c_n}) = \frac{\mathcal{L}_{\mathcal{R}}}{\int \mathcal{L}_{\mathcal{R}}\ ds} = \frac{P(\vec{c_n}|s)}{\int P(\vec{c_n}|s)ds} \tag{84}$$

The last equality in equation 84 is the same expression that "frequentists" use for calculating their errors after fitting, namely the likelihood curve normalized to unity gives the parameter errors. If the likelihood curve is Gaussian shaped, then this justifies a change of negative log-likelihood of $\frac{1}{2}$ from the optimum point to get the 1σ errors. Even if it is not Gaussian, as we show in section (10), we may use the expression for $P(s|\vec{c_n})$ as a *pdf* of the parameter s to evaluate the errors.

Note also that the expression for $P(s|\vec{c_n})$ in equation 84 is invariant under the co-ordinate transformation $c \rightarrow c'(c)$, since the Jacobian cancels in the numerator and denominator.

The normalization condition (using equation 75)

$$P^{data}(\vec{c_n}) = \int P(s, \vec{c_n})ds = \int P(c_n|s)\mathcal{P}_n(s_T)ds \tag{85}$$

is obeyed by our solution, since

$$\int P(\vec{c_n}|s)\mathcal{P}_n(s_T)\; ds = \int \alpha_n P(\vec{c_n}|s)\; ds \equiv P^{data}(\vec{c_n}) \tag{86}$$

The expression $\int \alpha_n P(\vec{c_n}|s)\; ds$ in the above equation may be thought of as being due to an "unknown concomitant" whose peak position is distributed uniformly in s space. The likelihoods of the theoretical prediction $P(\vec{c_n}|s)$ contribute with equal probability each with a weight α_n, to sum up to form the data likelihood $P^{data}(\vec{c_n})$. i.e. the data, due to its statistical inaccuracy will entertain a range of theoretical parameters. However, equation 86 does not give us any further information, since it is obeyed identically.

8.9 The dependence of α_n on n

For binned likelihood fitting, as $n \to \infty$, the likelihood ratio at $s = s_T$ will tend to $\exp(-n_b/2)$ where n_b is denotes the number of bins (see equation 36). We do not currently have an analytic theory for unbinned likelihood fitting. However, we can perhaps assume that the limit of the binned likelihood ratio approaches that of the unbinned likelihood ratio as $n_b \to \infty$ and $n \to \infty$. In either case then $\mathcal{L}_R(s_T)$ approaches a finite number ($\exp(-n_b/2)$ or 0). However, $P(s|\vec{c_n}) \to \delta(s - s_T)$ as $n \to \infty$. This must imply that $\alpha_n \to \infty$ in this limit, implying a dependence on n for $\mathcal{P}_n(s)$. This is another way of illustrating the difference between $\mathcal{P}_n(s)$ and the Bayesian prior, which is supposed to be a constant function, independent of n.

9 Combining Results of Experiments

Each experiment should publish a likelihood curve for its fit as well as a number for the data likelihood $P^{data}(\vec{c_n})$. Combining the results of two experiments with m and n experiments each, involves multiplying the likelihood ratios.

$$\mathcal{L_R}_{\ m+n}(s) = \mathcal{L_R}_{\ m}(s) \times \mathcal{L_R}_{\ n}(s) = \frac{P(\vec{c_m}|s)}{P^{data}(\vec{c_m})} \times \frac{P(\vec{c_n}|s)}{P^{data}(\vec{c_n})} \tag{87}$$

Posterior densities and goodness of fit can be deduced from the combined likelihood ratio.

10 Interpreting the results of one experiment

After performing a single experiment with n events, we now can calculate $P(s|\vec{c_n})$, using equation 84. Equation 78 gives the prescription for arriving at $\mathcal{P}_n(s)$, given an ensemble of such experiments. The ensemble is a purely theoretical abstraction. In practice, one only has a single dataset $\vec{c_n}$. If there were two such datasets, they would combined to form a single dataset $\vec{c_{2n}}$. One thus has to come to grips with interpreting the results of a single experiment. If the ensemble consists of N elements denoted by the index $k, k = 1, N$, then as $N \to \infty$,

$$\frac{dN}{N} \to P^{data}(\vec{c_n})d\vec{c_n} \tag{88}$$

The equation 78 can be written

$$\mathcal{P}_n(s) = \int P(s|\vec{c_n})P^{data}(\vec{c_n})d\vec{c_n} = \int P(s|\vec{c_n})\frac{dN}{N} \approx \frac{1}{N}\sum_{k=1}^{k=N} P(s|\vec{c_n}) \tag{89}$$

i.e.$\mathcal{P}_n(s)$ is the ensemble average of the posterior densities $P(s|\vec{c_n})$. Thus given a single experiment, the unbiased estimator for $\mathcal{P}_n(s)$, the *pdf* of s, is $P(s|\vec{c_n})$. We can thus use $P(s|\vec{c_n})$ as though it is the *pdf* of s and deduce limits and errors from it. The proviso is of course that these limits and errors as well as s^* come from a single experiment of finite statistics and as such are subject to statistical fluctuations.

11 Another Illustrative Example

We now apply the theory developed here to a practical example. The problem is to determine the weight of an object using an apparatus whose standard error is known to be 5 gm. The weight is a fixed constant of nature for the duration of the experiment. We obtain a dataset of 100 measurements, i.e. $n = 100$. Then $P(c|s)$ is a Gaussian of unknown mean s and width $\sigma =$ 5 gm. We compute $P(\vec{c_n}|s)$ for the 100 events by multiplying the individual $P(c_i|s)$ together and maximize the likelihood to determine s^* for the dataset using unbinned likelihoods. We then transform the measurements c_i to the hypercube space using equation 3. We use the improved PDE in hypercube space with $h = 0.2$ and determine the goodness of fit and the negative log-likelihood ratio $\mathcal{NLLR}$. We repeat this for an ensemble of 1000 experiments.

Figure 9(a) shows the distribution of s^* for this ensemble. The mean value of s^* over this ensemble is 49.98 gm and the RMS is 0.495 gm which is consistent with the expected $\sigma/\sqrt{(100)}$ value of 0.5 gm. Figure 9(b) shows the distribution of $\mathcal{NLLR}$ for the 1000 members of the ensemble. Figure 9(c) shows the likelihood ratio functions $\mathcal{L}_R(s)$ for the first 10 fits in the ensemble. The

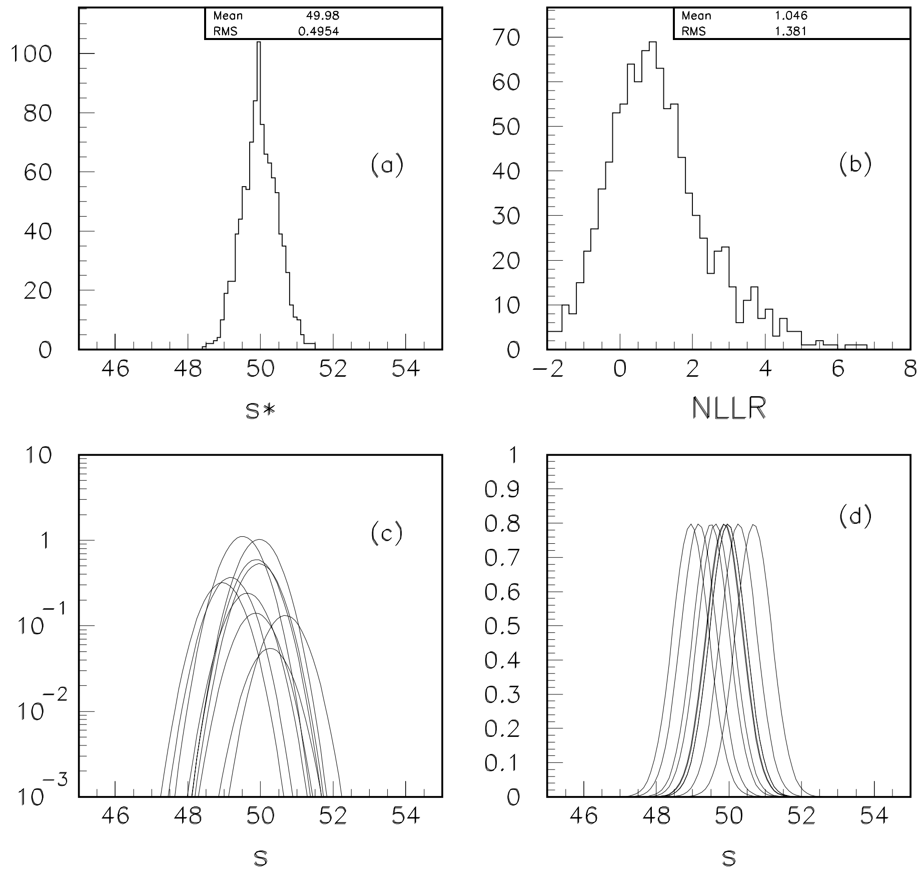

Fig. 9. (a)The distribution of s^*, the maximum likelihood value of s for a 1000 member ensemble of datasets of $n = 100$. (b)The goodness of fit variable $\mathcal{NLLR}$ for the fits (c)The likelihood ratio $\mathcal{L}_R(s)$ as a function of s for the first 10 members of the ensemble (d) The function $P(s|\vec{c_n})$ for the first 10 members of the ensemble

value of s^* fluctuates as expected, as well as the value of $\mathcal{L}_R(s^*)$, the negative logarithm of which gives the $\mathcal{NLLR}$. The fluctuation in s^* for the fits in the ensemble essentially expresses our lack of knowledge of the position of the true value s_T. The width of the likelihood distribution also contains information on the same lack of knowledge.

We now take each function $\mathcal{L}_R(s)$ and hypothesize that the true value is at a given value of s and apply Bayes' theorem as per equation 80. This set of infinite mutually exclusive hypotheses also expresses the same ignorance of the position of s_T. Bayes' theorem allows us to connect the two approaches (theoretical and data) to provide a calculation the posterior density $P(s|\vec{c_n})$ for each member of the ensemble. These functions are shown in Figure 9(d). The maximum likelihood value moves around with the expected spread of 0.5 gm. The average standard deviation of these curves is 0.5 gm with an rms of 0.65 E-3 gm. The average of these functions on an infinite ensemble yields the true *pdf* $\mathcal{P}_n(s)$.

11.1 One more iteration

In practice, if one has a dataset with $n = 100$ and $N = 1000$ similar instances of them, the easiest way to analyze the data is to combine them all into a dataset with $n' = Nn = 100,000$. However, we are interested in studying the function $\mathcal{P}_n(s)$ which is estimated by the ensemble average of the functions $P(s|\vec{c_n})$. This function tells us the behavior of the distribution of the maximum likelihood values s^* over similar datasets each with n=100.

After we do the average and obtain our best estimate of $\mathcal{P}_n(s)$ on the ensemble, we have more information (from the whole ensemble) on the position of the true value than we possessed while evaluating $P(s|\vec{c_n})$ for an element of the ensemble. We should use this additional information by re-introducing it into the Bayes' theorem equations 74 and 75 to re-work the individual $P(s|\vec{c_n})$.

$$P(s|\vec{c_n}) = \frac{P(\vec{c_n}|s)\mathcal{P}_n(s)}{\int P(\vec{c_n}|s)\mathcal{P}_n(s)ds} \tag{90}$$

where we approximate $\mathcal{P}_n(s)$ by the ensemble average. The resulting $P(s|\vec{c_n})$ are used to recompute the ensemble average to yield a better (iterated) estimate for $\mathcal{P}_n(s)$ as per equation 89. Figure 10(a) shows the ensemble average estimate of $\mathcal{P}_n(s)$ for n=100 and N=1000 before and after iteration. The mean value of the uniterated and iterated functions are the same at 49.977 gm (The Gaussians were generated with a true value of 50 gm). The r.m.s values of the function before and after iteration are 0.701 gm and 0.522 gm respectively. The iterated function thus has the correct width and mean value. Figure 10(b) shows the individual $P(s|\vec{c_n})$ functions for two members of the ensemble before and after iteration. The iterations pull these functions towards the true value, since we are inputing additional information on the true value.

12 The distribution of s^* and the function $\mathcal{P}_n(s)$

Using Bayes' theorem, we have shown that the function $\mathcal{P}_n(s)$ as estimated on the ensemble using equation 89 yields the *pdf* of s^*, the maximum likelihood values measured for each dataset on the ensemble. Here we show the same fact another way. The functions $P(s|\vec{c_n})$ are functions of s and depend on the individual dataset $\vec{c_n}$. Each dataset k in the ensemble yields two quantities after fitting and iteration; the maximum likelihood value s_k^* and the posterior density function $P(s|\vec{c_n})$. Without loss of generality, we can express the posterior density function as a function of $s - s_k^*$ such that

$$P(s|\vec{c_n}) \equiv \mathcal{G}_k(s - s_k^*) \tag{91}$$

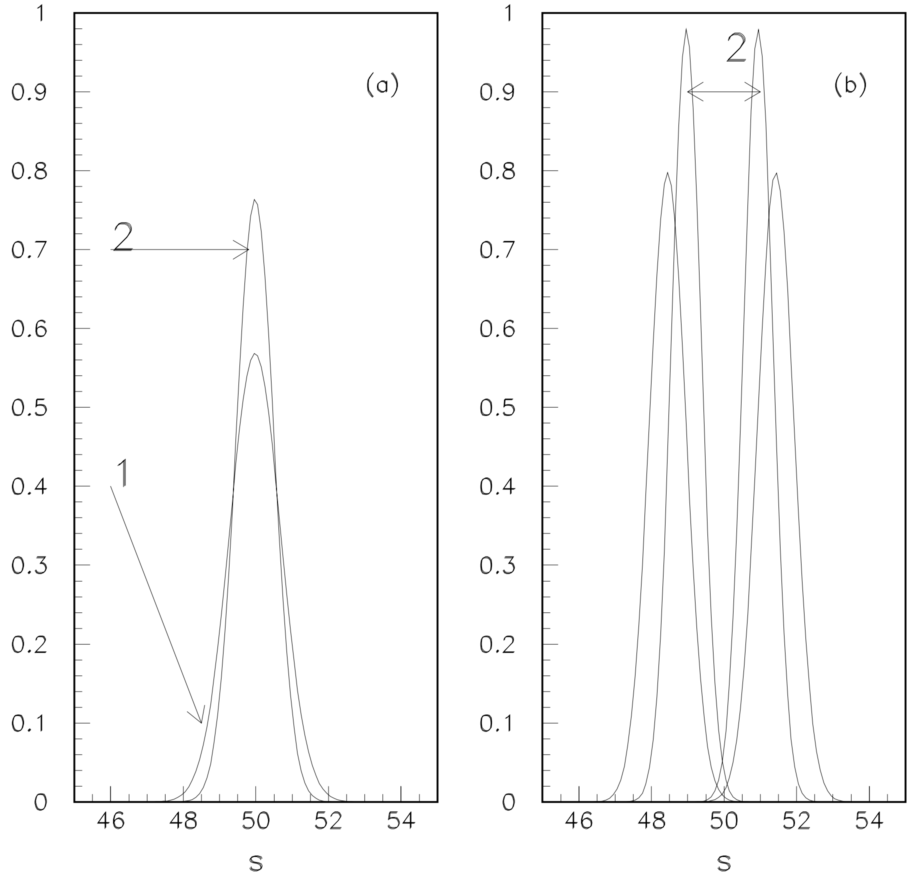

Fig. 10. (a) The function $\mathcal{P}_n(s)$ computed on the ensemble for n=100 and N=1000. The two iterations are shown, with the numbers (1,2) indicating the iteration number. (b) The function $P(s|\vec{c_n})$ for two elements on the ensemble for the two iterations. Then equation 89 can be re-expressed

$$\mathcal{P}_n(s) \approx \frac{1}{N} \sum_{k=1}^{k=N} \mathcal{G}_k(s - s_k^*) \tag{92}$$

But this is just the PDE equation for the distribution of s^*, with the functions $\mathcal{G}_k$ serving as the kernels!. They satisfy the normalization condition $\int \mathcal{G}_k(t)dt = 1$ as required. This should be compared with equation 50 for the definition of $PDE's$. Thus $\mathcal{P}_n(s)$ represents a PDE of the distribution of s^* and will yield the same distribution as s^*.

In the limit $N \to \infty$, we can represent the distribution of the maximum likelihood values s^* on the ensemble as a continuous *pdf* $g(s^*)$. In this limit, one can write

$$\mathcal{P}_n(s) = \int g(s^*)\mathcal{G}(s^*, s - s^*)ds^* = g(s) \tag{93}$$

where we have used the notation $\mathcal{G}(s^*, s-s^*)$ to emphasize the variation of the kernel as a function of s^* (i.e. ensemble element). The latter half of the above equation is an integral equation with kernel $\mathcal{G}(s^*, s - s^*)$ whose eigenfunction is $g(s)$. Figure 11(a) shows the values of s^* histogrammed for our illustrative example for an ensemble of N=1000 and n=100. The superimposed curve is the iterated function $\mathcal{P}_n(s)$ calculated for this ensemble normalized to a 1000 element ensemble. It can be seen that the function describes the distribution of s^* well. Figure 11(b) shows the iterated function $\mathcal{P}_n(s)$ for $n = 100$ and $n = 200$ respectively. As expected, the $n = 200$ function is narrower and its value at the maximum is larger, illustrating that $\alpha_n \equiv \mathcal{P}_n(s_T)$ increases with n.

13 Co-ordinate transformations $s' = s'(s)$

We have shown that the posterior densities $P(s|\vec{c_n})$ are invariant under the co-ordinate transformations $c' = c'(c)$, as they should be. How do they behave under transformations $s' = s'(s)$? The function $P(s|\vec{c_n})$ represents our estimate using one member of the ensemble of the *pdf* of s. So if $P(s|\vec{c_n})$ represents a

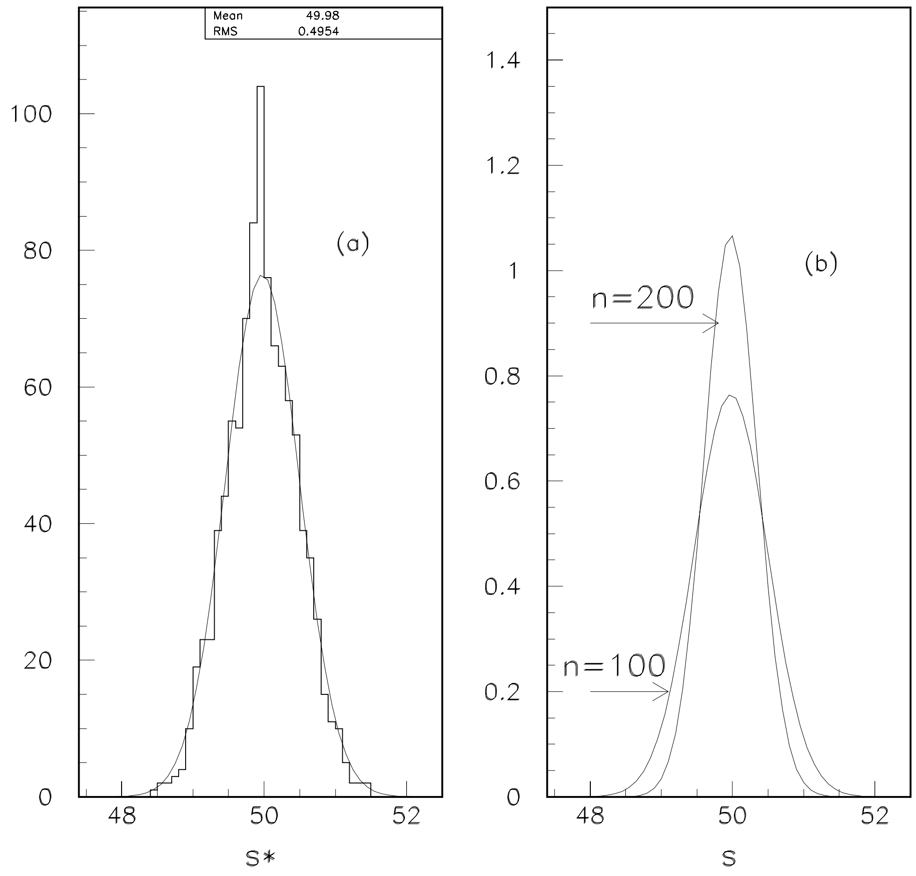

Fig. 11. (a)The distribution of s^* (solid histogram) for an ensemble with N=1000 elements each consisting of a dataset n=100. The curve is the estimate for the iterated function $\mathcal{P}_n(s)$ for this ensemble normalized to the 1000 observations. (b) $\mathcal{P}_n(s)$ on the ensemble for n=100 and n=200. This illustrates that the ensemble averaged function, depends on n, the size of the dataset. As n increases, the function narrows and the value of the function at its maximum increases.

pdf, we would expect it to behave like a *pdf*, namely

$$P(s'|\vec{c_n}) = P(s|\vec{c_n})|\frac{\partial s}{\partial s'}| \tag{94}$$

This is how $pdf's$ transform (via the Jacobian). This can be shown patently not to be so, since $P(\vec{c_n}|s') = P(\vec{c_n}|s)$ and

$$P(s'|\vec{c_n}) = \frac{P(\vec{c_n}|s')}{\int P(\vec{c_n}|s')ds'} = \lambda(\vec{c_n})P(s|\vec{c_n}) \tag{95}$$

where the s independent constant $\lambda(\vec{c_n})$ is given by

$$\lambda(\vec{c_n}) = \frac{\int P(\vec{c_n}|s)ds}{\int P(\vec{c_n}|s')ds'} \tag{96}$$

i.e. the posterior densities do not transform in a way that is expected of $pdf's$. This was perhaps a naive expectation. As we have just demonstrated, the posterior densities serve the purpose of kernels on the ensemble, the ensemble average of which gives the *pdf* $\mathcal{P}_n(s)$. There is no need for the kernel from a member of the ensemble to transform to the kernel from the same member under these transformations. The properties of the ensemble average deduced from the individual kernels will fluctuate from kernel to kernel. Similarly, when one analyzes in transformed variables, the same kernel will give different results which may be thought of as being part of the fluctuation.

The distributions of the maximum likelihoods $g(s^*)$ however will transform as $pdf's$, since $g(s)$ represents the probability density of the maximum likelihood value. i.e.

$$g'(s') = |\frac{\partial s}{\partial s'}|g(s) \tag{97}$$

Since we have demonstrated using equation 93 that $\mathcal{P}_n(s)$ and $g(s)$ are identical distributions, we can similarly assert that $\mathcal{P}'_n(s')$ and $g'(s')$ are identical distributions. And due to equation 97, we conclude that

$$\mathcal{P}'_n(s') = |\frac{\partial s}{\partial s'}|\mathcal{P}_n(s) \tag{98}$$

i.e the true $pdf's$ on an infinite ensemble will transform correctly. The individual kernels will not transform on to each other as $pdf's$.

14 Comparison with the Bayesian approach

In the Bayesian approach, an unknown Bayesian prior $P(s)$ is assumed for the distribution of the parameter s in the absence of any data. The shape of the prior is guessed at, based on subjective criteria or using other objective pieces of information. However, such a shape is not invariant under transformation of variables. For example, if we assume that the prior $P(s)$ is flat in s, then if we analyze the problem in s^2, it will not be flat in s^2. This feature of the Bayesian approach has caused controversy. Also, the notion of a *pdf* of the data does not exist and $P(c)$ is taken to be a normalization constant. As such, no goodness of fit criteria exist. In the method outlined here, we have used Bayes' theorem to calculate posterior densities of the fitted parameters while being able to compute the goodness of fit. The formalism developed here shows that what is conventionally thought of as a Bayesian prior distribution is in fact a normalization constant and what Bayesians think of as a normalization constant is in fact the *pdf* of the data. Table 2 outlines the major differences between the Bayesian approach and the new one.

15 Conclusions

To conclude, we have proposed a general theory for obtaining the goodness of fit in likelihood fits for both binned and unbinned likelihood fits.. In order to obtain a goodness of fit measure, one needs two likelihoods:- one derived

Table 2

The key points of difference between the Bayesian method and the new method.

Item	Bayesian Method	New Method
Goodness of fit	Absent	Now available in both binned and unbinned fits
Data	Used in evaluating theory *pdf* at data points	Used in evaluating theory *pdf* at data points as well as evaluating data *pdf* at data points
Prior	Is a distribution that is guessed based on "degrees of belief" Independent of data, monolithic	No prior needed. One calculates a constant from data $\alpha_n = \frac{P^{data}(\vec{c_n})}{\int P(\vec{c_n}\|s)ds}$ $\rightarrow \infty$ as $n \rightarrow \infty$
Posterior density $P(s\|\vec{c_n})$	Depends on Prior. $\frac{P(\vec{c_n}\|s)P(s)}{\int P(\vec{c_n}\|s)P(s)\ ds}$	Independent of prior. same as frequentists use $\frac{P(\vec{c_n}\|s)}{\int P(\vec{c_n}\|s)\ ds}$

from theory and the other derived from the data alone. In order to compute the errors on fitted quantities, posterior densities need to be worked out and Bayes' theorem needs to be employed. The usage of data likelihood using data alone does away the need for the Bayesian prior which is shown to be a number and not a distribution. This number is the value of the *pdf* of the parameter, which we call the "unknown concomitant" at the true value of the parameter. This number is calculated from a combination of data and theory and is seen to be an irrelevant parameter. If this viewpoint is accepted, the controversial practice of guessing distributions for the "Bayesian Prior" can now be abandoned, as can be the terms "Bayesian" and "frequentist". We investigate the transformation properties of the posterior density of fitted parameters under change of variable.

16 Acknowledgments

This work is supported by Department of Energy.

References

[1] R. A. Fisher,"On the mathematical foundations of theoretical statistics", *Philos. Trans. R. Soc. London Ser. A* **222**, 309-368(1922);
R. A. Fisher,"Theory of statistical estimation", *Proc. Cambridge Philos. Soc.* **22**, 700-725 (1925).

[2] E. Parzen, "On estimation of a probability density function and mode" *Ann.Math.Statis.* **32**, 1065-1072 (1962).

[3] S. Baker, R. D. Cousins, Nucl. Instrum. Meth. A221 (1984).

[4] "A measure of the goodness of fit in unbinned likelihood fits", R .Raja, Fermilab-PUB-02-152-E, physics/0207083

[5] D. Scott. *Multivariate Density Estimation.* John Wiley & Sons, 1992.
M. Wand and M. Jones, *Kernel Smoothing.* Chapman & Hall, 1995.

[6] "An essay towards solving a problem in the doctrine of chances", Rev. Thomas Bayes, Biometrika,**45** 293-315 (Reprint of 1763) (1958).